REPORT

on the

USE OF MEDI-CAL MANAGED CARE

ENCOUNTER DATA FOR RESEARCH PURPOSES

January, 2002

Prepared by

Jim Klein
Medical Care Statistics Section
California Department of Health Services
714 P St., Rm 1750
Sacramento, CA 95814
Telephone: (916) 657-0893
Email: jklein@dhs.ca.gov

GRAY DAVIS
Governor
State of California

Grantland Johnson
Secretary
California Health and Human Services Agency

Diana M. Bontá, R.N., Dr.P.H.
Director
Department of Health Services

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The author gratefully acknowledges The California Endowment and California Wellness Foundation for jointly providing the funding necessary to carry out this study.

I also gratefully acknowledge the invaluable guidance and inexhaustible encouragement of Dr. Larry Dickey, who has played a key role in all phases of this project.

Several people with a depth of knowledge and experience on Medicaid encounter data issues provided countless insights and practical ideas, and patiently answered my many questions on the subject. The list is long, but it includes most prominently Lisa Chimento and Joel Menges of The Lewin Group, Kathe Fox of The MEDSTAT Group, and Nelda McCall of Laguna Research Associates. Insights about encounter data and how they are captured by managed care plans were contributed by Aimee Chitayat of the Medi-Cal Policy Institute.

Significant contributions to the section on deliveries were made by my manager, Gene Hiehle, and Penny Cook, a consultant on this project. Wes Ingram, of the Medi-Cal Managed Care Division, provided countless edits and technical improvements on many parts of the report.

LIST OF FIGURES AND TABLES

Figures

Tables

Executive Summary

A study of the quality and completeness of California's Medicaid medical managed care encounter data was conducted during calendar year 2000 by the Medical Care Statistics Section of the California Department of Health Services (DHS) with funding from the California Endowment and California Wellness Foundations. To properly structure this review and to gain an understanding of the best uses of managed care encounter data, a review of the literature on administrative databases was performed and experts on the collection of encounter data were consulted.

Largely as a means to ensure greater access to care, in 1994 DHS began transitioning a large segment of the Medicaid (Medi-Cal) population from a fee-for-service (FFS) medical delivery system to a managed care delivery system. In 1994, 86% of the Medi-Cal population received their services through FFS, while in 2001 over 50% receive their medical services through managed care plans. Almost 80% of those on Temporary Assistance to Needy Families (CalWorks in California) are now covered by managed care. With these changes, DHS and other researchers are interested in knowing what the effects of managed care are on the medical care and health status for this population. One of the primary means for determining these effects is to study the encounter data submitted by the managed care plans for documentation of medical services rendered to plan enrollees.

For over 15 years, claims data have been compiled into computerized administrative databases for Medicaid, Medicare, and Veterans Administration medical services. These electronic fee-for-service claims data have been used with much success for administrative and health services research purposes including: setting reimbursement rates, developing more effective protocols for health care, improving access to care, monitoring the emergence of new diseases, and identifying over- and underutilization of care by certain populations. Accessing these databases is relatively easy and inexpensive, and sampling bias when studying specific areas is avoided. As administrative databases were more extensively studied, their shortcomings became more evident. Claims data accurately reflect some kinds of information, such as identity of recipients, while other types of information are not as accurately captured, such as disease diagnosis. Administrative data cannot be used to determine beneficiaries' pre-treatment states of health, lab results, most outcome conditions (e.g., birth weight, mortality, quality of life), and care obtained outside the medical delivery program collecting the claims data (e.g., free clinics, services outside the scope of benefits).

Despite these shortcomings, the medical industry, particularly managed care organizations, is becoming increasingly reliant on administrative encounter data. Factors promoting these trends include: improvements in and greater pervasiveness of information technology tools; competitive pressures in the managed care arena; and impending implementation of the Health Insurance

Portability and Accountability Act, which will standardize coding schemes in the health care industry.

In this project, five indicators were utilized to assess the quality of Medi-Cal encounter data: timeliness, validity, completeness, accuracy, and consistency. Ten-percent samples of fee-for-service databases for the two periods of fiscal year (FY) 1994-95 and calendar year (CY) 1999 were created to serve as benchmarks for these indicators. In preparation for the databases, duplicates were removed. The encounter data were found to have significant rates of duplicate records (9.0%, 7.8%, 6.3%, and 3.6% for the first through fourth quarters of CY99, respectively).

Timeliness. Managed care data were found not to be submitted in as timely a manner as FFS data. The approximate number of days from the date of service of encounter claims to the date the record enters DHS claims files ranged from 183 days to 98 days from the first to the last quarter of CY99. By comparison, the average lag for FFS data in 1999 was 43 days.

Validity. With one notable exception, Medi-Cal encounter data were found to contain valid entries. Validity tests conducted on the encounter data resulted in very low error rates for diagnosis codes (0.8%), procedure codes (0.8%), procedure modifiers (1.5%), national drug codes (0.69%), provider types (except for one plan, error-free), and physician specialty (0.6%). Some plans had much higher rates of invalid values, but usually less than 10%. A high rate of validity was found for beneficiary identification numbers in the form of the MEDS-ID (usually Social Security Number). These were verified against eligibility records for the correct month and plan assignment 97.9% of the time for all plans; eliminating one plan with a high error rate increased the rate for the other plans to 98.7%.

Most encounter data records were found to contain provider numbers that could not be validated, either because they were not listed in files maintained by DHS or because DHS does not have access to a master list. Only 16.0% of the records had a provider number that could be matched to the Medi-Cal provider master file (28% non-drug, 3.5% drug), and 6.5% had a provider number in the form of a license number that was matched to this provider master file. An additional 4.8% of the provider numbers were matched to lists provided by the managed care plans or to the license number file provided by the Department of Consumer Affairs. Other forms of the provider identifiers permitted for use by Two-Plan/Geographic Managed Care (GMC) plans, such as facility, tax and national identification numbers, could not be checked for validity. When the information on provider name, provider type, and zip code given on the encounter record was matched to the same information on the provider master file, the successful match rates varied from 34% to 92% by health plan. Information for these three encounter data variables (provider name, provider type, zip code) was evaluated to determine if they were unique to the provider number given by each of the plans in their encounter data. On average, for each

unique provider number given by a plan, there were 1.8 unique names, 1.1 unique provider types, and 1.3 unique zip codes. The inability to identify most providers rendering services to managed care recipients may seriously undermine the use of encounter data for such purposes as monitoring access to providers with certain specialties or cultural/linguistic characteristics.

Accuracy. Managed care encounter data were compared to FFS data in order to evaluate accuracy for two issues: the relative frequency of the general therapeutic class of prescribed drugs, and the relative frequency of the asthma diagnostic codes. In both cases little difference was found between managed care and FFS data, indicating that for these two categories of data, managed care data may be as accurate as FFS data.

Completeness. Managed care data were found to be very incomplete in comparison to FFS data. Encounter data completeness was reviewed by comparing the rate of encounters per thousand beneficiaries to FFS data for FY94-95 and CY99, with counts expressed as a rate of encounters per thousand beneficiaries. Because some suggest that managed care plans successfully report at least one encounter per visit but not necessarily the second or third due to the paperwork burden, a comparison was performed using the number of encounter days per beneficiary. The results were substantially the same whether the FY94-95 FFS or CY99 FFS data base was used, or encounters or encounter days were used as a unit of measure. Encounter data had about 33% of the FFS encounters for inpatient, 38% of the medical/outpatient encounters, and 76% of the drug encounters. (Because under managed care the days supply per prescription may be approximately double for many drugs than under FFS [30 vs. 60 days], the 76% drug completion factor may, in effect, be closer to 40%.) A separate estimate of the completion rate for inpatient encounters was made for deliveries, and the result correlated well with the overall inpatient completion rate.

Consistency. Health plans submitting fewer encounters per thousand beneficiaries tended to be less consistent in the submission of their data when measured over time. The encounter data were reviewed for consistency of submission on a month-by-month basis by looking at the three lowest encounter volume months versus the three highest volume months. Recognizing that demand for medical services is higher some months due to seasonal variation, e.g., allergy and flu seasons, the overall utilization should be approximately the same month to month. For encounter data, the lowest three months by plan only had about 50% of the encounters as the three highest months, suggesting that there are gaps in the data submitted by the plans.

In summary, Medi-Cal managed care encounter data for CY99 were substantially incomplete. Provider identification numbers can seldom be used to identify unique providers by name and location. For the other critical data elements, the managed care data were found to be reasonably valid. Timeliness of data submission is probably not a significant problem for managed care data,

although submission is delayed relative to FFS data. Given the relative strengths and weaknesses of Medi-Cal managed care data, many researchers may find that it requires too much effort to query and analyze it properly. On the other hand, researchers could begin to gain some valuable insights into what is occurring in Medi-Cal managed care by looking at only those plans with substantially complete data. Only with increased analysis and use of encounter data will its full strengths and weaknesses be known, and strategies developed for further improvements.

I. Introduction

The DHS Research Policy Council met during 1998 and 1999 to help define priority research issues related to the transition of the state's Medi-Cal program from a fee-for-service (FFS) model to a predominantly managed care model. A basic issue of concern that emerged from these meetings was the difficulty faced by researchers in accessing Medi-Cal data for studies. To address this issue, two foundations (The California Endowment and the California Wellness Foundation) provided grants ($22,000 each) to DHS through the Office of Clinical Preventive Medicine (OCPM), which had convened the DHS Research Policy Council.

A new Data Access Work Group was convened for this project consisting of: Chief, Office of Clinical Preventive Medicine (chair); Chief, Medi-Cal Managed Care Division, Chief; Medical Care Statistics Section, Chief, Management Information Systems/Decision Support Systems; and Chief, Medi-Cal Policy Division. Funds were provided to the Medical Care Statistics Section (MCSS) to carry out the analytical work for the project. The MCSS has the responsibility within DHS for ad hoc analysis of FFS Medi-Cal medical claims data. These analyses are used within State government for budgeting, program administration, and policy research. The MCSS also routinely releases FFS claims data to the public in encrypted formats for academic and public program policy research purposes. Due to the complexities inherent in the Medi-Cal program and its data files, MCSS has developed manuals and guides for DHS and non-DHS data users that define and explain these data and their interrelationships, and how the data relate to the program.

This project, as defined by the Data Access Work Group, has two major components. The first is addressed in this report: evaluating the quality of Medi-Cal managed care encounter data for research purposes. Because of concern about encounter data quality, the Data Access Work Group decided that encounter data should not be released until this work could be completed. The second component, not reported here, was to clarify administrative procedures and lines of authority for data release to outside researchers.

What is Encounter Data?

Encounter data reflect activity at the level of the provider-patient interaction. What diagnosis was made? What tests were performed? What treatments were provided? These are a few of the important issues that encounter data can address. In FFS delivery systems, much of this information can be obtained from billing claims submitted to the payer of services, such as Medi-Cal. However, in capitated delivery systems, such as Medi-Cal managed care, payment is not tied to specific encounters, but to insured patients. This disconnect of payment from encounter removes a major motivation of providers to submit complete, reliable information. Because of the potential value of encounter data for evaluating health care access and quality of care issues,

Medicaid managed care plans are contractually required to submit encounter data from providers.

Table 1, on the next page, provides a matrix of potential uses and stakeholders of Medi-Cal managed care data. The following table, Table 2, provides a matrix of key data elements required for each use. The key data elements are critical not only on an individual encounter record basis, but are needed in linking many encounters either within the same database, or to information in external sources, such as the Office of Statewide Health Planning and Development's hospital discharge files, and vital statistics birth files. In addition, beneficiary IDs that are consistently and accurately coded can be used for random sampling from the encounter database.

Table 1.
Matrix of Uses and Stakeholders of Managed Care Data

General Purpose	Use	Primary Stakeholder				
		DHS	Plans	Private Drug/ Medical Firms	Medical/ Managed Care Researchers	Medi-Cal Beneficiaries
Health Services Research	Quality of care, e.g., HEDIS*	X	X	X	X	
	Effectiveness of health protocols	X	X	X	X	
	Effectiveness of health technologies and drugs	X	X	X	X	
	Effectiveness/efficiency of managed care vs. FFS	X			X	
	Illness prevalence			X	X	
	Variations in care by geography, ethnicity, etc.	X			X	X
	Episodes of care	X		X	X	X
	Outcomes	X	X		X	X
	Access to care (e.g., to specialists, to linguistically-appropriate providers)	X	X		X	X
Administrative	Effectiveness of contracted plans	X	X			
	Efficiency of contracted plans	X	X			
	Individual provider assessment	X	X			X
	Plan stability/performance/auditing/contract monitoring	X	X			
	Program change analysis, e.g., benefits, eligibility	X				
	Rate setting	X	X			

*Health Plan Employer Data and Information Set

Table 2.
Matrix of Uses and Data Elements of Managed Care Data

General Purpose	Use	Data Elements							
		Required						Preferred	
		Diag-nosis Code	Procedure/ Drug Code	BeneficiaryID	Date of Service	Provider/ Claim Type	Provider ID	Hlth Care Plan	Aid Code
Health Services Research	Quality of care, e.g., HEDIS*	X	X	X	X	X			
	Effectiveness of health protocols	X	X	X	X	X	X		
	Effectiveness of health technologies and drugs	X	X	X	X	X			
	Effectiveness/efficiency of managed care vs. FFS	X	X	X	X	X		X	
	Illness prevalence	X	X	X	X	X			
	Variations in care by geography, ethnicity, etc.	X	X	X	X	X			
	Episodes of care	X	X	X	X	X			
	Outcomes	X	X	X	X	X	X		
	Access to care (e.g., to specialists, to linguistic-ally-appropriate doctor)			X	X	X	X	X	X
Administrative	Effectiveness of contracted plans	X	X	X		X		X	
	Efficiency of contracted plans	X	X	X		X		X	
	Individual provider assessment	X	X	X	X	X	X		
	Plan stability/performance/auditing/contract monitoring	X	X	X	X	X		X	
	Program change analysis, e.g., benefits, eligibility	X	X	X	X	X			X
	Rate setting			X	X	X		X	X

*Health Plan Employer Data and Information Set

Submission and Processing of Medi-Cal Encounter Data

Managed care encounter data are submitted electronically in different formats and to different entities, depending on the type of managed care organization (MCO) submitting them. For Two-Plan and Geographic Managed Care (GMC) plans, encounter data must be received by the State within 90 days of date of service. There is no similar requirement for the County Organized Health Systems (COHS) plans. COHS plans submit their encounter data electronically directly to DHS in a file format known as the "35-File." This file format is mainframe-compatible, variable-blocked, in an EBCDIC character set, and includes data formats unique to a mainframe environment, such as binary and packed-decimal. Since 1988, when the San Mateo and Santa Barbara COHS plans started submitting data to the State, the task of reviewing the files has been that of the MCSS and the Information Technology Services Division (ITSD). These reviews have largely consisted of detecting major problems, especially those that might cause "fatal" processing errors. A detailed and structured review of COHS data for completeness and quality has never been undertaken.

Two-Plan and GMC plans submit data directly to the Medi-Cal fiscal intermediary, Electronic Data Systems (EDS) using electronic media, e.g., cartridges, tapes, and online. The files are edited by EDS for three types of errors: "critical error," "one-percent error," and "five-percent error." "Critical errors" which cause a file to be rejected include: no header record present; invalid submitter ID code; record count on transmittal does not match record count from file; and any single record has incorrect values for the data elements of plan code, claim type (e.g., medical), segment counts for inpatient encounters, or adjudication status code. "One- and five-percent error" editing involves determining the percent of validity errors for specified data elements and then, if the one- or five-percent error threshold is reached on any of these data elements, the whole file is rejected. For instance, provider type is a "one-percent error" data element. Thus, if a file had more than one percent of its provider type values that were invalid (i.e., not in the table of acceptable provider type values), the whole file is rejected. "One-percent error" threshold data elements include: those that must have values specified in the data element manual (provider type, long term care accommodation, and ancillary codes); must be numeric (paid amount, Medicare deductible and coinsurance amounts, and days stay); or must meet other relational edits (beginning date of service must be earlier than ending date of service, and vice-versa). "Five-percent errors" include numeric checks (billed amount and procedure quantity), checking of values contained in external Medi-Cal files (beneficiary Medi-Cal number, birth date, National Drug Code), and other checks.

Encounter data files that pass these checks are sent each month to the ITSD, which reformats them for the 35-File. These records are combined with those from the COHS and are sent to the Management Information System/Decision Support System maintained by The MEDSTAT Group.

To better determine how the process of collecting managed care data may affect its quality and completeness, the Medi-Cal Policy Institute commissioned Outlook Associates to assess the Medi-Cal medical managed care data flow and processes from the time a beneficiary enters a provider's office until those data reach the State. The report, "From Provider to Policymaker: The Rocky Path of Medi-Cal Managed Care Data," (March 2001) found: 1) there was a lack of collaboration and communication among DHS and its vendors in the collection of encounter data;

2) there was inadequate focus on data accuracy, completeness and timeliness; 3) there was a lack of standardization in the Medi-Cal managed care information system; and, 4) current information systems and processes at all levels of the system have limited capacity to respond to changing needs.

II. Medicaid Managed Care Encounter Data Quality: Review of Current Knowledge

Medicaid programs in a number of other states were contacted regarding their experiences with collecting and analyzing managed care encounter data. Only a few states believe that their data are approaching the completeness and quality of FFS data. These include Arizona, Colorado, Maryland, Oregon, Tennessee, and West Virginia. Two of these (Oregon and Tennessee) have gone so far as to use their encounter data for setting managed care capitation rates, even to the extent of applying risk adjustment methodologies for some of the beneficiary population. Most states, however, are struggling to obtain good quality encounter data and to use those data in program administration and quality improvement efforts. Barriers reported by states include: 1) fewer managed care organizations bidding for Medicaid business; 2) a lack of an information technology (IT) infrastructure in the health care industry, particularly traditional Medicaid safety net providers; and, 3) lack of highly trained IT staff at the state level overseeing and maintaining the collection and use of these data.

To understand how managed care encounter data may be useful for research and administrative purposes, a perspective on how medical administrative data sets have been used in the past is necessary. Administrative data in the form of electronic media can be extremely valuable, but there also are many shortcomings to their use. Awareness of all the strengths and weaknesses of this type of information will help ensure their use to attain realistic objectives.

Advantages of Administrative Data

FFS delivery systems have been instrumental in advancing the creation and use of administrative databases since the reporting of the details about the encounter was necessary for providers of care to get reimbursed. Administrative databases created through Medicare, Medicaid, the Department of Veterans Affairs, and the Canadian health care system, as well as state hospital discharge files, have been particularly valuable to medical researchers in evaluating issues of quality of care and access issues. Other users of these types of databases have included clinicians, consumers, government, providers, Health Maintenance Organizations (HMOs), and insurance companies.

Administrative databases containing medical encounter or claims information have many advantages over other forms of medical experience information. Garnick, et al. evaluated the value of such databases, and enumerated some of the advantages:

> Currently available administrative datasets can offer a number of advantages for studies of quality of care compared with data from medical records, surveys or special studies:
>
> - Individuals can be tracked over time (potentially both before and after treatment).
> - Enough patients are included to allow studies of specific providers.

- Large populations make it possible to select enough people with relatively rare conditions.
- Services can be identified regardless of site of care or provider.
- Researchers do not influence practice through their data collection (i.e., no Hawthorne effect).
- Bias is not introduced by including only high quality clinical centers.
- Research costs are lower than for the collection of primary data.[1]

Other researchers have cited additional advantages of administrative databases. Among them are these cited by Welch and Welch: setting capitation rates, monitoring underuse, providing medical information to HMOs, helping Medicare manage its FFS sector, exploring the relationship between utilization and health, and investigating changes in medical practice and incidence of disease.[2] Other researchers have found similar benefits in using administrative data sets. [3, 4, 5]

Researchers have used administrative databases extensively over the last several years to great effect. However, as their uses have expanded, so have the cautions about drawing too many conclusions from these types of data, or expecting administrative data to answer questions they are inadequate to answer. Databases with claims submitted under a FFS arrangement and edited by a claims processor will contain more accurate and complete information than those submitted without such financial incentives and/or without extensive automated checking. On the other hand, claims submitted under a FFS reimbursement arrangement will have a higher rate of duplication (to the extent edits will not detect them), "upcoding" of diagnosis and procedure codes (under which services are billed at a higher reimbursement level than they should be), and fraud and abuse. [6]

Limitations of Administrative Data

One limitation of administrative data sets is the inaccuracy of information for some variables. Generally, the variables that are more likely to be edited by the claims processor, or which do not depend on a high degree of judgment by the claims coder, are more likely to be accurate. Thus, beneficiary and provider identifier numbers are often correct because they must match with records in the enrollment and provider files at the claims processor. Likewise, values for gender, beneficiary birth date/age[7], and date of service [8] are generally accurate because the correct information is readily known and available, and thus obvious errors are easy to identify by the claims coder and the claims processor. Diagnosis and procedure codes are relatively inaccurately reported, with diagnosis being the least accurately reported. Schwartz, et al. provide an enumeration of the reasons for this:

> Often administrative data are weak in the identification of cases using diagnosis codes for a number of reasons: physician unfamiliarity with the use of ICD-9-CM coding and therefore a lack of congruity between the charting by the physician and the abstraction of the principal diagnosis by medical staff; ICD-9-CM codes often do not allow for adequate discrimination among cases with the same disease because the diagnosis alone does not capture the severity; physician listing diagnoses on the chart that are suspected but not necessarily final; and upcoding of severity to enhance payment.[7]

Researchers at RAND have emphasized the imprecision of ICD-9 coding:

> The lack of operational definitions for ICD-9-CM diagnostic codes tends to result in highly variable assignment of codes to health conditions. From a data user's perspective, this makes it difficult to extract all relevant cases within a particular diagnosis code category to examine questions about the costs or patterns of care for a particular condition. For example, bronchitis and upper respiratory illnesses may be coded using different ICD-9-CM codes but may refer to the same condition. . . . Diagnosis codes tend to encompass broad ranges of disease severity and therefore mask important clinical subgroups that differ in their expected response to treatment. Also, diagnosis codes do not allow direct determination of the patient's severity of illness . . . [which] is important to control for . . . .[9]

In studies comparing individual claims in the administrative data set with other sources, e.g., the medical chart, the ICD-9 is found to be miscoded 65 to 80% of the time.[10, 11, 12] When the level of specificity was limited to the first three digits of the ICD-9 code, accuracy was over 90%.

Procedure codes have been found to be relatively accurately coded. This is attributable in large part to the greater specificity within the Current Procedure Terminology (CPT) coding scheme. Some suggest, however, that it may also be attributable to medical practitioners' greater knowledge about entering procedure codes than diagnosis codes. Other critical data elements, such as provider number, place of service, quantity, and claim type, have not been as extensively studied.

Generally, data will be more accurately and completely coded when:

- The coder has fewer choices of values to code.
- Incorrect coding is readily apparent or codes are more frequently used, e.g., incorrect dates are more readily detected than incorrect beneficiary Social Security Numbers (SSNs); breast cancer is more accurately coded than the rare condition of Cushing's Disease.
- Coded information is derived from automated and linked sources, e.g., gender from eligibility files are more accurately coded than diagnosis codes.
- Claims coding practices within the medical office or plan have been in place for an extended period and thus would result in stable and consistent coding practices.
- There are negative consequences to the provider when submitting bad information, e.g., the claims will be rejected for payment or will be paid at a lower level.

With regard to the latter, Weiner, et al. have noted: "In general, the greater the importance to a payer or payee, the higher the likelihood that corresponding data items will be captured by the claims system with reliability and completeness . . . . Clearly, when data are needed to set payment levels, incentives exist for both parties – payer and provider – to ensure that data are captured accurately."[8] The provision of small financial incentives to managed care providers to submit accurate and complete automated encounter data is often not sufficient to ensure the submission of accurate data.[13] Others have expressed caution about using administrative data submitted under capitated arrangements unless significant validation efforts using such sources as medical charts are employed.[14] Unfortunately, the medical record itself is scarcely the gold standard that many believe: in a recent study it was found that the medical chart included

mention of services in the chart that were not provided (false-positives) at a rate of 19%, and did not include services that were performed (false-negatives) at a rate of 30%. It was also found that the rate at which an incorrect diagnosis was recorded was significant.[15]

Iezzoni has succinctly noted: "Administrative data cannot elucidate the interpersonal quality of care, evaluate the technical quality of processes of care, determine most errors of omission or commission, or assess appropriateness of care." [16] In addition to these weaknesses, Poses et al. point out the reasons why it is so difficult using administrative data to gather pretreatment prognostic characteristics about patients to make a valid correction for prognostic differences among patients receiving different treatments:

> Existing administrative databases may contain sociodemographic information and some information about comorbidity . . . . They may not capture other important kinds of prognostic information that may affect physicians' decisions, the most important of which is the severity of a specific presenting problem, but including specific co-morbid problems that may be particularly relevant in the context of a specific presenting problem; the severity of co-morbid diseases; baseline health status; and biopsychosocial factors.[4]

These and other shortcomings limit the usefulness of claims data; such other shortcomings include:

- lack of documentation of laboratory and diagnostic test results, education, counseling and preventive services;
- itemized services billed under global codes, e.g., prenatal services, immunizations;
- services outside the scope of benefits of the insurer, e.g., those from free clinics, those paid for directly by the beneficiary within the deductible amount; and
- the timing of the assignment of a diagnostic code that may influence the decision on the appropriateness of care or severity of illness.[8, 9, 13, 17, 18]

In summary, the experience with medical administrative data sets over the last two decades has shown that, even under the best conditions, there are many and substantial limitations restricting their use. The inherent variability in data accuracy and their sensitivity to disparate factors in their preparation and submittal must be expected when assessing the integrity of administrative data, and their limitations must similarly be recognized when setting realistic objectives for use of managed care encounter data.

The Future of Administrative Encounter Data

Despite these shortcomings, the medical industry is moving toward documenting services rendered by managed care organizations using administrative data. Several factors are responsible for this. First, the use of information technology within the medical industry is inexorably increasing, and concomitantly, database software and hardware standards are converging to a common set of standards. The electronic medical chart may yet be years away

from general use or acceptance, but the transcription of rendered goods and services into a general health database is becoming more common.

Second, competitiveness in the managed care health services sector has led health care organizations to use administrative information to become more efficient and effective, and to promote their companies to the consumer through such means as health plan report cards. Also, using administrative data to set risk-adjusted capitation rates and establish model protocols of care may help health care organizations to become more efficient and effective, while delivering a higher quality product to their enrollees. Some analysts maintain that the aggressive use of information technology may very well determine which organizations survive in the competitive health care environment.

Third, the impending implementation of the Health Insurance Portability and Accountability Act to standardize coding of electronic medical transaction data will encourage use of standardized coding schemes throughout the health industry and encourage electronic capture of health encounter transactions.

Fourth, the Centers for Medicare and Medicaid Services ((CMS), formerly the Health Care Financing Administration) implemented provisions of the Balanced Budget Act of 1997, which requires reporting of both FFS and managed care encounters for all Medicaid-covered services. These services must be reported to CMS through the submission of data to the Medicaid Statistical Information System (MSIS). CMS is becoming involved in actively monitoring the completeness and quality of these data, having contracted with The MEDSTAT Group to evaluate the completeness and accuracy of the states' Medicaid encounter data submitted to CMS for their MSIS data base. In addition, under this project, known as "The Evaluation of Medicaid Encounter Data Project," MEDSTAT will also develop a straightforward and simple process for evaluating encounter data in the future so that it will be able to use these data to support critical required analyses. Data from each of the 30 states submitting data will be evaluated by comparing each state to others as well as doing internal reviews of each state. Specific objectives of the project are to:

- identify reports and other information that States are generating from encounter data in order to provide examples of reporting that HCFA may wish to consider for replication, normative reporting, etc.;
- evaluate and report MSIS encounter data completeness by reviewing standard MSIS data quality reports and creating supplemental data completeness reports; and
- develop and conduct basic analyses regarding the use of encounter data to support Medicaid managed care program management and oversight.

In a related project, the Center for Health Care Strategies has contracted with Mathematica Policy Research and, as a subcontractor, The Lewin Group, to survey the states on the general uses for their Medicaid managed care encounter data, especially for rate-setting. Their report is expected to be released early in 2002.

The quality of Medi-Cal managed care encounter data should improve due to the efforts of the DHS. DHS' Management Information System/Decision Support System Office has implemented a Statistical Process Control (SPC) system that calculates measures of the

encounter data completeness across several variables, and plans to share SPC results with the plans to help them identify gaps in their encounter data. The Medi-Cal Managed Care Division (MMCD) is working with managed care plans to improve the completeness and quality of their provider cross-reference files, and is coordinating with ITSD to develop an automated process for checking data submitted by county organized health system plans. MMCD is also participating in the effort to replace the current file-level encounter data rejection system with a record-level rejection system. The Medi-Cal fiscal intermediary, EDS, is implementing this change, and expects to begin during 2002. This would affect all Two-Plan and GMC data currently processed by EDS, as well as data from one COHS.

III. Statistical Analysis of Medi-Cal Managed Care Encounter Data

Establishment of the Databases

The first step in the statistical review was the creation of three databases: an historical FFS pre-managed care database for FY 94-95, a contemporaneous FFS database for CY99, and a managed care encounter database for CY99. All three databases were limited to mandatory aid codes for the Two-Plan and GMC plans (0A, 01, 02, 08, 3A, 3C, 3E, 3G, 3H, 3L, 3M, 3N, 3P, 3R, 3U, 30, 32, 33, 34, 35, 38, 39, 5X, 54, 59, 82). Twenty-one months of "paid" or processed data were used to create a managed care database with 12 months of incurred (service date) data. The characteristics of the three databases are shown in Table 3.

Table 3.
Characteristics of Study Databases

Type of Coverage	Time Period	Counties Covered	Sample Size
FFS	Jul-94 - Jun95	Alameda, Contra Costa, Fresno, Kern, Los Angeles, Monterey, Orange, Riverside, San Bernardino, San Diego, San Francisco, San Joaquin, Santa Clara, Santa Cruz, Stanislaus, Tulare	10%
FFS	CY99	All claims for persons not in a managed care plan	10%
Managed Care	CY99	Same counties as for July, 1994-June, 1995 plus Napa, Sacramento San Mateo, Santa Barbara, Solano	100%

All managed care encounter data for CY99 were utilized because the volume was manageable without sampling. The FFS databases were 10% samples because utilizing all FFS data would have been unmanageable due to volume. The 10% samples were established by

selecting all records having a pair of digits in last part of the beneficiary's SSN that matched one of ten pairs of random numbers. Experience in using such sample files over the last ten years has shown this to be a valid and reliable sampling method.

Next, claim adjustments were deleted from the databases. Out of the 21 processed months used for the managed care database, the months of December 1999, January 1999, April 2000, and July 2000 had the highest rates of adjustments: 12%, 45%, 37%, and 8%, respectively. The other 17 months had an average of 1.4% adjustments of all records.

When selecting the 12 months of incurred records, an initial check was performed on the "date of service" data for managed care. (The dates of service from the FFS data were assumed to be valid due to edits they are subject to at the fiscal intermediary.) Out of 26 million records in the encounter database, less than 100 were found with either invalid date formats for the date field, or dates greater than the month in which the data were processed by the State. It should be noted that the dates of service fields for inpatient records for Two-Plan/GMC records were blank. This was found to be due to the fact that these fields were not required in the claim format for inpatient/Two-Plan-GMC encounters. For these records, dates of admission and discharge were used as proxies for "from" and "to" dates of service.

To ensure comparability, encounter data were subjected to the same duplicate editing process used for FFS data. In this process a claim was tagged as a duplicate if it was not an adjustment and the values for the following fields were the same as for a previous claim: beneficiary ID, provider number, procedure code and modifier or National Drug Code, "from" date of service, "to" date of service, and units. This differed from DHS' current duplicate detection methodology for encounter data in that it additionally used the claim reference numbers (CRN) field as a criterion. CRNs are issued by the health plan for each submitted encounter record. This method thus assumes that the health plans are able to detect duplicate encounters submitted by their providers before issuing the number. Although duplicates were removed from the FFS databases for this study, apparent duplicate claims could remain due to reinstatement upon manual review by claims suspense examiners. Because the managed care encounter database could not be reviewed on a record-by-record basis for these exceptions, all apparent duplicates were removed from the FFS database. The rate of duplicates for the FFS FY94-95 database was less than one percent (0.72%). The duplicate rates by quarter for the managed care encounter data are shown in Table 4.

Table 4.
Duplicate Encounter Record Rates by Quarter

Calendar Quarter of Service	Duplicate Rate
Jan.-Mar.	9.0%
Apr.-Jun.	7.8%
Jul.-Sep.	6.3%
Oct.-Dec.	3.6%

Researchers should be aware that Medi-Cal encounter data may contain duplicates not detected by the CRN editing methodology, and DHS should consider monitoring these duplicates in an ongoing fashion using the same duplicate editing methodology used for FFS data.

Characteristics of Data Quality

Five characteristics were utilized to study the quality of Medi-Cal encounter data: timeliness, completeness, accuracy, validity, and consistency. These characteristics are widely used in the field to evaluate the overall usefulness and integrity of administrative data bases. [9, 15, 19, 20, 21]

Timeliness is how quickly data become available for use by the State after the date of service. This period incorporates the time it takes for the encounter record to be sent to the State, and the time it takes the State (including the fiscal intermediary) to process that record before it becomes available. Timeliness is critical to the usability of the data for such purposes as setting rates and quickly implementing revised medical protocols to maximize their effectiveness and efficiency.

Validity is the degree to which data in each field use acceptable values or formats. Acceptable values or formats may be defined in referenced tables, subsidiary files, or categorization schemes. The most direct and straightforward method for verifying data validity in claims records is to check the values in each and every field against either the list of possible values or formats prescribed in the data element dictionary. A more sophisticated validity test, not utilized in this study, involves checking the value in one field against that in another using a "validity map;" this type of checking is also referred to as "relationship checking" or cross-field editing. For example, only certain procedure codes can be used with certain diagnosis codes. Similarly, providers of one type are licensed to provide only certain types of goods or services, e.g., an optometrist cannot perform surgery.

Accuracy is the degree to which the provision of medical goods and services is correctly documented in the data. For example, if the patient receives a sigmoidoscopy but the provider documents a mammogram, the data are inaccurate. An inaccurately coded field on the encounter record can result in either a valid or an invalid value for that variable. An analysis of the frequency of logical groupings of values by variable can help reveal if the inaccuracy was due to misjudgment on the part of the coder, or to carelessness. In the case of the former, most of the codes will be valid but the profile of the distribution of values will not match typical distributions for that variable. In the case of carelessness, there will be more invalid values present in the frequency distribution. Methods of checking coding accuracy against external sources are similar to those mentioned above for checking data completeness.

Completeness is the degree to which the data capture each and every good or service rendered to enrolled beneficiaries. The best way to determine completeness of a data set is to compare the electronic encounter record to documentation of the encounter kept at the provider's location, such as the medical chart. Because such field audits are time-consuming and expensive, this method was outside the scope of this study. However, proxy completeness standards developed from studying FFS claims were used

for this project to determine the degree of encounter data completeness. Completeness of the data can be assessed by making internal comparisons within the database, or by comparing data against external sources. Both approaches were used for this project.

One approach not within the scope of this study, but feasible when the encounter data is reasonably complete, is to determine situations when more than one encounter would be necessary for a single event, then identify if both types of encounters appear in the record. Such paired-encounters for single events might include diabetes and insulin, or asthma and inhalants. Before analyzing the encounter database for completeness and other factors, claims for services or goods rendered more than once, i.e., duplicate claims, must be removed from both the FFS and encounter data databases.

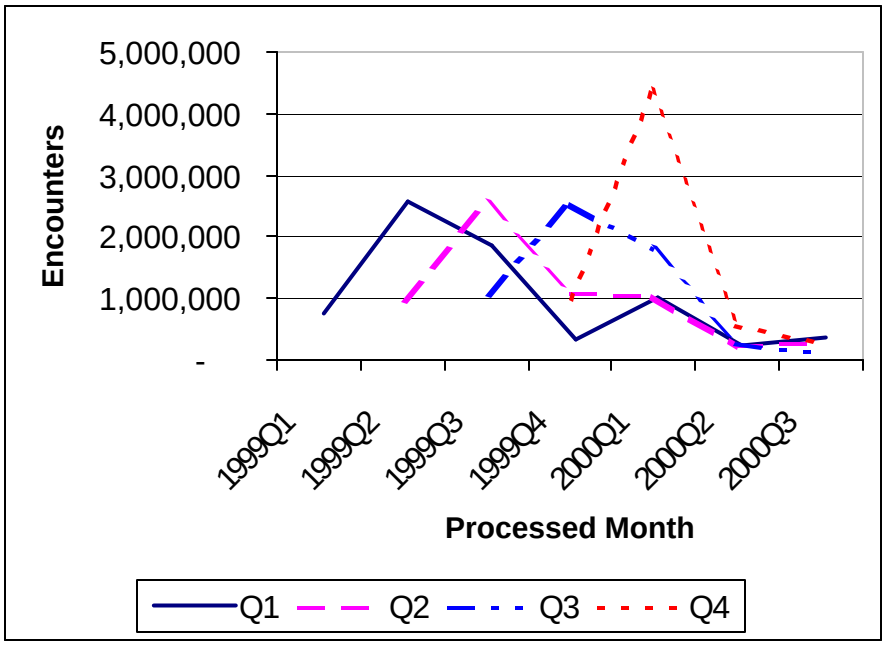
Consistency is the degree to which data are similar across time in content and volume. Inconsistent data is much more difficult to statistically adjust than data with consistent, known data-quality weaknesses. Inconsistent volumes of data being reported over time would suggest either under- or over-reporting since beneficiary populations in aggregate have approximately the same level-of-service demand over time. Unexpected variability across plans may reflect differences in interpretations of the State's data requirements, or the strengths or weaknesses of different MIS systems. Inconsistencies within one plan's set of encounter data may be useful in identifying weaknesses in that plan's underlying manual and automated information systems that capture and process the data. In general, as a health care organization matures in capturing and reporting encounter data, the data should become more consistent.

The following discussion of encounter data quality is organized according to these five basic characteristics. However, the reader should keep in mind that, although these qualities are distinguishable from one another conceptually, in practice they are difficult to measure independently. For instance, completeness can only be tested when it is known that the data were submitted in a timely manner. Likewise, data values can be both inaccurate and invalid. It is nonetheless helpful to review encounter data using each of these five characteristics since this scheme adds structure to the analysis and may pinpoint causes of poor data quality that may readily lead to correction of problems.

Timeliness of Encounter Data

To estimate the time lag between date of service to the date that the State has finished processing the data, the number of records by service date quarter was graphed against the quarter they were processed by the State, as shown in Figure 1. (Note: Q1 = Jan-Mar99, etc.)

Figure 1.
Encounters per Service Data Calendar
vs. Processed Month Calendar Quarter



To obtain a more precise estimate of how long it takes for data to become available from date of service to when the State can access it for administrative uses, the average lag time was calculated from month of service to the month the encounter was added to DHS encounter tapes. The results are shown in Table 5.

Table 5.
Average Lag Time in Days and Percent of Records
By Non-COHS vs. COHS Plans,
And Calendar Quarter

	All Records		**Non-COHS**		**COHS**	
CY99 Date of Service Quarter	Days Lag to Processed Month	% of Records	Days Lag to Processed Month	% of Records	Days Lag to Processed Month	% of Records
Jan.-Mar.	183	27.6%	192	27.7%	111	26.5%
Apr.-Jun.	150	24.5%	153	24.5%	126	24.7%
Jul.-Sep.	114	23.1%	114	23.2%	116	22.6%
Oct.-Dec.	98	24.8%	99	24.6%	91	26.2%
Average	**136**	**100.0%**	**140**	**100.0%**	**111**	**100.0%**

In general, encounter data are available within 4 months of service delivery, a period short enough to allow health plans or others to make interventions and track potential outcomes

closely. There was significant improvement over time in the lag from date of service to the month the data were processed by the State for the non-COHS (i.e., Two-Plan and GMC) plans. This improvement was not as great for the COHS plans, except for the fourth quarter. Because there does not appear to be a significant correlation between the days of lag by service quarter and the percent of the records for that quarter, the increased number of processed months data "available" to the first quarter versus the last quarter did not seem to make much difference in the lag periods. An alternative explanation for the differences in lag periods across quarters seems to be a genuine improvement in the timeliness in the submission of encounter records by the plans.

Validity of Encounter Data

For this portion of the study CY99 data for the key data elements listed in Table 2 were tested against acceptable values listed either in the Managed Care Data Dictionary or DHS' 35-File Manual, as shown below in Table 6 below.

Table 6.
Data Elements and Tables of Valid Values

Data Element	Table of Valid Values
ICD9, 3- and 5-bytes	Electronic Data Systems Files/ ICD9-CM manual
Procedure Code	Electronic Data Systems Files/ CPT/HCPCS manuals
Procedure Modifier	Modifier lists, various
National Drug Code (NDC)	Electronic Data Systems Files/ First Data Bank
Provider Vendor Code	List of specified DHS values
Provider type	List of specified DHS values
Provider specialty	List of specified DHS values
Claim type	List of specified DHS values
Health Care Plan (HCP) number	List of specified DHS values

ICD9 Diagnosis Codes Validity. The first three and five bytes of ICD9 values for both the primary and, when present, the secondary fields were examined for all non-drug managed care encounters. At the three-byte level, the rate of invalid codes for the primary diagnosis code was 0.028%, and the rate for the secondary diagnosis code was 0.024%. The three highest invalid rates for primary diagnoses for managed care plans were 15%, 10% and 5%. The COHS plans had a lower overall rate than the Two-Plan/GMC plans. For provider types with more than one thousand encounters, the rural health/federally qualified health clinics had the highest invalid rate for primary diagnoses, at 0.9%.

At the five-byte level, the invalid rate was 0.84% for the primary ICD9 code, and 0.23% for the secondary code. The three highest rates for primary diagnoses for the managed care plans

were 50.2%, 6.5%, and 1.1%. It should be noted that the primary reason that one plan had a 50% invalid rate was that the fifth byte was consistently filled in with a zero or a comma. Similar to the three-byte diagnosis code, the provider type with more than 1,000 encounters with the highest invalid diagnosis rate was the rural health/federally qualified health clinics, at 1.0%, followed by community inpatient hospitals, at 0.7%.

Procedure Codes Validity. The procedure code values for medical and outpatient claims were assessed using the current list of procedure codes for FFS claims (Current Procedural Terminology (CPT)-4 and HCFA Common Procedure Coding System (HCPCS) codes). Procedure codes '00001' through '00022' are valid for rural health/federally qualified health clinics for the COHS', but not the Two-Plan/GMC plans. Taking these differences into account, the invalid rate for all health plans was 0.78%. Outpatient claims had a slightly higher rate (0.88%) than medical services (0.77%). The three highest invalid rates for health plans were 5.9%, 5.4%, and 4.5%. The provider types with the highest invalid rates were community clinics (3.4%), rural health/federally qualified health clinics (2.8%), and community outpatient hospitals (1.0%).

Procedure Modifier Codes Validity. Procedure modifiers are used to provide more specific information about procedures than procedure codes themselves permit. Some procedure codes must have a modifier while others may or may not, depending on the service. The procedure modifier code values for medical and outpatient claims were reviewed for valid values using two sets of modifiers (one for COHS plans and one for the Two-Plan/GMC plans) specified in the Encounter Data Element Dictionary. It should be noted that these two lists of modifiers have substantial differences. Of the 89 modifiers that are valid for the COHS plans, 20 are not contained in the Two-Plan/GMC list. Conversely, of the 167 in the Two-Plan/GMC list, 97 are not on the COHS list. Another difference is that the Two-Plan/GMC list includes 36 modifiers that can be used only when medical transportation codes are used.

Taking these differences into account, 1.52% of the modifiers were invalid in the encounter data. Notably, the outpatient claim type had an invalid rate of only 0.02%, whereas the medical services claim type had an invalid rate of 1.78%. There were a significant number of invalid modifier values of '00' for some plans. When these were disregarded, one plan had an invalid rate of 5.6%, one with 3.9%, and 10 with invalid rates from 1.0 to 2.0%. The provider types with the highest percent of invalid modifiers were: miscellaneous (9.3%), physicians (4.3%), physician groups (4.0%), and optometrists (3.2%).

Data users should be cognizant of these coding differences. If appropriate, DHS may want to consider using the same modifier coding scheme used in FFS to avoid misinterpretations of modifiers in encounter data for Two-Plan/GMC plans versus those from COHS's.

National Drug and Supply Code Validity. The NDC and Medi-Cal supply codes were also checked for validity. It should be noted that the Two-Plan/GMC data element dictionary allows for one supply code not listed for use by the COHS plans, '9999M-ZZ' for miscellaneous supplies. Taking this into account, the invalid rate for NDC codes was 0.69%, and for Medi-Cal supply codes it was 0.14%. The highest invalid rate by plan was 19%, with six plans having an invalid rate of between one and five percent. DHS should assess the feasibility of disallowing supply code '9999M-ZZ' inasmuch as it is not a valid code for the COHS plans.

Provider Type Validity. Virtually all the provider type values were valid in the encounter claims, the only exception being a COHS plan that did not furnish this information to the State. Although the value of '99' is permitted to designate "miscellaneous" provider types, high use of this code may indicate that the use of this value is being abused. The highest rates at which '99' was used by plans were 21%, 20%, 9%, 4%, 1.7%, and 1.2%. In contrast, for the CY99 FFS data base, there were no claims with a provider type coded as '99.'

Physician Specialty Validity. Physician specialty values are required whenever the encounter record indicates that the provider is a physician or physician group. The overall rate of invalid values for encounter data was 0.6%. The rate of miscellaneous and unknown values was 4.0%. One plan had an invalid rate of 30%, and four plans had rates between one and two percent.

Managed Care Plan and Claim Type Validity. Because managed care plan codes are validated before encounter data are accepted by the State, all these codes were valid. Claim type codes, indicating outpatient, inpatient, drug, and medical, were found to be substantially correct throughout the encounter data base.

Provider Number Validity. The requirements for provider number values are different for the COHS than the Two-Plan/GMC plans. Except for pharmacies, COHS plans must submit encounter data with an approved Medi-Cal provider number. The Two-Plan/GMC plans can submit the Medi-Cal provider number (if assigned), or, if not assigned, the provider's State license number. Clinics can submit the State clinic license number. The Two Plan/GMC plans may also provide either their own plan-assigned provider identification number, a tax identification number, or a national provider identification number.

Two approaches were taken to validate the provider numbers submitted on encounter claims: 1) a match of provider number known from external provider files, and 2) a check of the uniqueness of each provider number within and across plans using other fields on the encounter record, i.e., name, zip code, and provider type.

Three "external" provider files were used to validate the provider numbers on encounter claims: the Medi-Cal Provider Master File (PMF), the Plan Provider file (built with provider submissions from the plans), and a license number file produced by the Department of Consumer Affairs. The PMF was used not only to match against the Medi-Cal provider number, but also to match for the state license number. The rates at which the provider number from the encounter records matched to various provider files is provided in Table 7.

Table 7.
Validity Match Rates Overall and by Claim Type of Provider Numbers

	Overall	Outpatient	Inpatient	Drug	Medical
Medi-Cal PMF.	16%	27%	33%	4%	28%
License numbers in PMF	7%	<1%	0%	<1%	15%
Prov Lists -- Plans'	4%	21%	21%	<1%	4%
License number file	1%	0%	0%	<1%	2%
No Match	73%	53%	46%	95%	52%

It is notable that the match rates by Medi-Cal provider number for drug claims was only 4%, significantly lower than the other claim types; the no-match rate was 95%, considerably higher than for the other claim types. The reason for this is under investigation.

For the encounter records with a provider number that matched to the PMF, the information provided on the encounter record for each provider was compared to the same information associated with that provider on the PMF. For these, the value for provider type matched 78% of the time, the first three bytes of the zip code matched 82% of the time, and the last name matched 65%. The rates for a successful match to the state license number on the PMF was 92% for provider type, 68% for first-3 bytes of the zip code, and 34% for the last name.

It should be noted that the low rate of match against the license number file is probably due in large part to the fact that the value in the license file is 12 bytes, while the number received on the encounter data is only 9 bytes. Most 12-byte license numbers have three or more embedded zeros, but the algorithm for removing these zeros varies. (Different algorithms were employed here to maximize the chances for a match.)

To determine if the managed care plans were consistently using unique information for each of their provider identification numbers, a count of unique provider identification numbers was compared to a count of unique provider names, provider types, and zip codes within each of the plans. Unique counts across all plans were also obtained. The results are shown in Table 8.

Table 8.
Ratios of Unique Provider IDs to Values for Unique Provider Names, Unique Provider Types and Unique Provider Zip Codes, Within and Across Plans

	Average, Within Plan	Across All Plans
Unique Names	1.81	1.28
Unique Provider Types	1.10	1.13
Unique Zip Codes	1.27	1.67

The ability to uniquely identify providers is important for several DHS monitoring functions:

- Monitoring ratios of providers to beneficiaries

- Monitoring type/specialty and geographical distribution of providers
- Monitoring how well a plan is meeting the cultural/linguistic needs of enrolled beneficiaries
- Monitoring the overlap of provider networks across plans
- Monitoring the use of traditional safety-net providers by plans
- Monitoring fraud and abuse.

Because of the lack of uniqueness of provider identification numbers on encounter claims vis-à-vis provider type, names, and zip codes, providers cannot be tracked uniquely. The many different provider number formats used by Two-Plan/GMC plans cannot be validated due to lack of access to the master file for each of these. Additionally, some allowable provider identifiers are tied to individuals (license number) or organizations (facility and tax IDs). However, the health plans are not constrained to use such identifiers for the individual, organization, or address to which the number was officially assigned. A potential solution to this problem would be for health plans to provide DHS with a list of all provider numbers and associated unique provider name (person or organization, as appropriate to the provider type it is tied to), provider type, specialty, address, and languages/cultures in which they are competent. DHS could use this information to perform regular checks of the provider identifiers in the encounter data against master provider files.

Beneficiary Number Validity. The tests on beneficiary eligibility entailed looking at both the identifier on the encounter record as well as other information on the record applying to that beneficiary (such as birth date, gender, county, aid code). Two-Plan/GMC plans must submit either the Medi-Cal Eligibility number (MEDS-ID), usually the SSN, or the Client Index Number (CIN). COHS plans must submit the MEDS-ID or the Beneficiary Identification Number. The fiscal intermediary, EDS, reviews Two-Plan/GMC encounter data and passes the MEDS-ID to the State, for inclusion in the 35-File. (This file was used for this project and would likely be used for release of encounter data to researchers outside DHS.)

When attempting to verify beneficiary MEDS-IDs in the encounter data (35-File) files, a significant percent of the values in the MEDS-ID field were found to be CIN numbers. The percent of CIN numbers in this field ranged from zero to 2.2% for the months of January through September 1999, then abruptly increased to 9% in October and eventually 53% in March 2000. This problem was reported to those involved with processing these files, and is being researched as to the cause. Related to this problem, a significant number of CIN numbers were found to have been submitted by a large health care plan for several months, except that the CINs from this plan were invalid by virtue of the fact that one of the nine bytes was reported in error in a consistent manner.

When verifying eligibility, the records with a beneficiary identifier number in the CIN format were first omitted. Verification of eligibility occurred when the MEDS-ID on the encounter record matched the State's eligibility file, the month of service matched the month of eligibility, and the plan reporting the encounter matched the plan that the beneficiary was actively enrolled in for that month.

Table 9 shows the beneficiary identifier error rates both including and excluding data for the one plan with the large number of errors cited above.

Table 9.
Beneficiary Identifier Error (Non-Match) Rates at the Plan Level

Month	Error Rate, All Plans	Error Rate, Excluding Outlier Plan
Jan-99	1.8%	1.4%
Feb-99	1.7%	1.3%
Mar-99	1.7%	1.3%
Apr-99	1.6%	1.2%
May-99	1.5%	1.0%
Jun-99	1.6%	1.0%
Jul-99	1.8%	1.2%
Aug-99	2.0%	1.3%
Sep-99	2.7%	1.8%
Oct-99	3.3%	1.9%
Nov-99	3.3%	1.8%
Dec-99	3.3%	1.6%
Total for CY99	2.1%	1.3%

Those plans with a mismatch rate of 2% or higher accounted for 36% of the encounters. With exclusion of the one plan with the very high error rate, plans mismatch rates of 2% or higher accounted for 17% of the encounters.

For beneficiaries with MEDS-ID verified eligibility, the values for birth date, gender, county and aid code was compared between the encounter record and the eligibility record. The rates of mismatch on these values are shown in Table 10.

Table 10.
Error (Non-Match) Rates for Beneficiary Information Comparing DHS Eligibility File to Encounter Record Values

Variable	CY99	Jan-Jun 99	Jul-Dec 99
Birth Date	1.3%	1.6%	0.8%
Gender	1.1%	1.3%	0.6%
County	0.2%	0.2%	0.1%
Aid Code	7.9%	17.6%	18.5%

The rates of mismatch were quite low for all but aid code. The rates (except for aid code) for the second half of CY99 were about half of those for the first half. This could be due to improvements by the plans in the accuracy of submitted encounter data, or related to the CIN problem referred to above. The mismatch rates for aid code were found to be due to the fact that the eligibility file used for checking these encounter data were built several months after the month of eligibility to account for retroactivity transactions. In examining the mismatches on aid code, virtually all could be accounted for due to retroactive changes in aid code, especially within the "30's" aid code series. For instance, often a beneficiary was initially assigned to aid code 30 or 38, then reassigned months later to the other. These changes would have no significant effect on coverage provided by or capitation rates paid to the managed care plan.

Accuracy of Encounter Data

Accuracy of the Medi-Cal encounter data for the national drug code (NDC) was evaluated by performing frequency comparisons to FFS data using both the FY94-95 and CY99 data sets. To facilitate the comparison, the general therapeutic class (GTC) associated with the NDC on the encounter/claim records was determined, then the respective distributions of percentages that each GTC contributed to all drugs was determined. Only the twenty highest GTCs by encounter/claim volume were used. The results using this method are shown in Table 11.

Table 11.
Comparison of Drug Records for Top Twenty GTC Percentages
FFS vs. Managed Care

GTC Description	FFS 94-95	FFS 99	Managed Care
Antiinfectives	24.0%	19.1%	22.0%
Cough/Cold Preparations	15.8%	11.3%	11.3%
Analgesics	12.1%	10.7%	10.1%
Skin Preps	8.1%	7.3%	6.5%
Eent* Preps	3.3%	3.4%	7.9%
Antiarthritics	2.8%	5.6%	6.9%
Antiasthmatics	4.7%	5.5%	4.9%
Gastrointestinal	4.3%	4.1%	3.2%
Antiinfectives/Miscellaneous	1.7%	2.4%	3.7%
Psychotherapeutic Drugs	2.6%	4.6%	3.0%
Hormones	2.0%	3.3%	2.9%
Elect/Caloric/H2O	4.4%	3.6%	1.8%
Antihistamines	2.1%	4.1%	2.2%
Vitamins	3.3%	1.6%	1.9%
Contraceptives	1.3%	1.4%	2.1%
Cardiovascular	0.9%	1.8%	2.0%
Unlisted	1.9%	1.3%	1.1%
Hypoglycemics	0.8%	1.8%	1.2%
Cardiac Drugs	0.8%	0.9%	0.8%
Autonomic Drugs	0.5%	0.8%	0.6%

*EENT: Eye, Ear, Nose, Throat

The managed care encounter drug data is similar to FFS data from either period in terms of types of drugs and their relative volumes being prescribed. This is especially notable given the several ways that patterns of care reflected in the beneficiary population represented here could be different (rural vs. urban; FFS vs. managed care; patterns in 1994 vs. 1999). This would seem to suggest that the accuracy and/or completeness of the coding of NDC numbers is relatively the same as FFS. Future research might be directed toward understanding the greatest differences between managed care and FFS relative to utilization of specific types of drugs, especially those for which the percentage of managed care encounters are significantly different from FFS claims.

Accuracy of the Medi-Cal encounter data was also analyzed by studying the frequency of diagnostic codes for asthma. Asthma is a condition whose rates should not be significantly affected by delivery model since it is an illness that cannot be cured, only controlled. Because of possible geographic factors affecting asthma, CY99 encounter data were compared to FY94-95 FFS data, rather than CY99 FFS data (which is heavily weighted with rural counties). The percentage of encounter data with an asthma diagnosis code (2.25%) is close to that of FFS for

FY94-95 (2.11%). This suggests that the coding of diagnosis groups may be substantially accurate under managed care.

Completeness of Encounter Data

To assess the completeness of encounter data, the number of encounters reported by managed care plans was compared to those reported by FFS providers, expressed as a count per thousand beneficiary months. Because the average number of encounters/claims varies depending on the type of service, the analysis included the four claim types: outpatient, inpatient, drug, and medical.

Table 12.
Count (per 1000 beneficiary months) and Rate of Encounters for FFS vs. Managed Care, By Claim Type

	FFS FY94-95	FFS CY99	<----------Managed Care---------->		
	<------ Count of Encounters ------->			% FFS 94-95	% FFS 99
Drug	606	518	462	76.3%	89.3%
Inpatient	9	7	3	33.3%	42.9%
Medical	857	497	394	46.0%	79.2%
Outpatient	357	708	70	19.6%	9.9%
All	1,884	1,785	950	50.4%	53.2%
Medical and Outpatient	1,214	1,205	464	38.2%	38.5%

The results in Table 12 suggest that overall the encounter data is about 50% complete, compared to either the FFS FY94-95 (50.4%) or the FFS CY99 (53.2%) databases. It is interesting to note that the rate of outpatient vs. medical claims per thousand beneficiaries was quite different between FY94-95 and CY99. The FFS94-95 rate has two and one-half times as many medical claims (857) as outpatient claims (357), whereas the CY99 baseline has about 70% as many outpatient as medical claims. This may be explained by the fact that the Medi-Cal population in non-urban (non-managed care) areas receive most of their care from physicians instead of clinic or outpatient facilities, whereas the opposite is true in the urban areas. In fact, if the claims per thousand beneficiary months for outpatient and medical are combined, the difference between these two periods is negligible (1214 vs. 1205).

It has been suggested that managed care plan providers may not generate as many encounters for a given visit compared to fee-for-service providers due to the paperwork burden. To explore this possibility, the encounter data rates from the previous section were re-calculated on the basis of encounter days per thousand beneficiary months rather than encounters per thousand beneficiaries. If the health plans were documenting every visit to a provider but generating fewer encounters per visit, the completion factors for managed care would be expected to increase relative to FFS using this analysis. Table 13 provides the results of this re-analysis.

Table 13.
Count and Rate of Encounter Days for FFS vs. Managed Care,
And by Claim Type

	FFS FY94-95	FFS CY99	<----------Managed Care---------->		
	<------ Count of Encounters ------->			% FFS 94-95	% FFS 99
Drug	305	308	229	75.1%	74.4%
Inpatient	9	7	3	33.3%	42.9%
Medical	456	282	214	46.9%	75.9%
Outpatient	161	403	25	15.5%	6.2%
All	931	1,000	471	50.6%	47.1%
Medical and Outpatient	617	685	239	38.7%	35.0%

The relative distribution of counts for managed care encounter data relative to FFS data is not significantly changed by this re-analysis, suggesting that managed care providers are generating as many claims per visit as FFS providers, but not documenting as many visits.

The completeness of encounters was found to vary considerably by health plan. The rate of encounters per 1000 beneficiary months by plan and claim type is reported in Table 14, below.

Table 14.
Rate of Encounters for FFS vs. Managed Care, by Claim Type and Plan

Plan Name	Plan No.	Outpat	Inpat	Drug	Medical	All	W/o drug
FFS 99	--	**708**	**62**	**518**	**497**	**1785**	**1267**
FFS 94/95	--	**357**	**65**	**606**	**857**	**1884**	**1279**
Blue Cross/Stanislaus	310	278	47	1137	716	2178	1040
Blue Cross/Kern	342	188	52	987	766	1992	1005
Blue Cross/Fresno	341	186	45	990	710	1931	941
Blue Cross/San Diego	048	168	46	859	846	1920	1060
Blue Cross/Sacramento	190	142	55	794	704	1695	901
Blue Cross/Tulare	311	196	45	782	574	1598	815
Blue Cross/Santa Clara	345	60	23	1112	353	1547	435
Blue Cross/Alameda	340	176	50	676	645	1547	871
Blue Cross/San Francisco	343	199	33	788	485	1505	717
Blue Cross/Contra Costa	344	193	39	680	467	1379	698
Omni/Stanislaus	359	215	47	519	593	1373	855
Health Plan Of San Mateo	503	439	69	364	498	1370	1006
Health Net/Fresno	351	239	39	460	619	1357	897
Health Net /Tulare	353	190	24	532	592	1338	806
Partnership Healthplan/Napa	507	102	47	399	750	1297	898
Contra Costa Health Plan	301	77	44	446	713	1281	835
Kern Health Systems	303	8	43	455	770	1276	821
Santa Barbara Health Initiative	502	40	36	453	743	1272	819
Blue Cross/San Joaquin	358	179	50	483	496	1209	726
Partnership Healthplan/Solano	504	54	56	350	696	1156	807
Central Coast Alliance/Santa Cruz	505	232	68	345	509	1154	809
CalOptima/Orange	506	71	26	666	377	1139	473
Blue Cross/Sacramento	180	109	36	348	594	1087	740
Inland Empire Health Plan (San Bernardino)	306	151	38	440	433	1061	621
Alameda Alliance For Health	300	1	34	437	573	1044	607
Central Coast Alliance/Monterey	508	249	73	301	387	1011	710
Health Plan Of San Joaquin	308	143	40	535	259	977	443
OVERALL AVERAGE		**70**	**24**	**462**	**394**	**950**	**487**
Inland Empire Health Plan	305	80	31	405	400	917	511
Health Net/Sacramento	150	63	17	396	397	873	477
Western Health Advantage	140	143	50	56	566	815	759
Health Net/San Diego	068	115	16	377	239	747	370
Health Net/Los Angeles	352	16	2	437	272	726	289
Kaiser/Sacramento	170	0	1	42	634	678	636
UC San Diego Health Plan	049	6	9	314	336	665	351
LA Care Health Plan	304	39	15	287	305	647	359
Community Hlth Grp/San Diego	029	0	26	521	61	608	87
San Francisco Health Plan	307	17	0	426	55	497	72
Santa Clara Family Health Plan	309	34	1	222	232	489	267
Universal Care/San Diego	023	28	2	378	40	449	71
Molina/San Bernardino	356	0	22	294	61	377	83
Molina/Riverside	355	0	16	245	52	313	68
Maxicare/Sacramento	160	0	15	34	8	58	23

To further investigate completeness of the encounter data, one type of service (deliveries) was reviewed using specific diagnosis codes. A methodology was developed to derive an estimate of expected deliveries by managed care plan. First, the number of FFS deliveries per 1,000 eligible months in the mandatory aid codes for CY99 was calculated for each ethnicity statewide and by county. The rates, standard deviations and coefficients of variation are shown in Table 15.

Table 15.
Statistical Measures of Deliveries for FFS Medi-Cal Beneficiaries

	Hispanic	White	Asian	Black
Eligible Months	287,652	602,825	53,811	26,835
No. of Counties*	14	34	10	7
Average Rate/County**	10.3	7.4	8.0	8.1
Standard Deviation	1.6	1.2	1.2	2.0
Coefficient of Variation	0.15	0.16	0.16	0.25
Statewide Rate	10.4	7.4	8.1	8.1

* Counties with 2,000 or more eligible months

** Delivery rates per 1,000 eligible months for women 15-44 years in non-managed care counties.

The coefficients of variation (CVs) provided a means for comparing the standard deviations across groups with substantially different average rates. The fact that the CVs are small suggests that ethnicity-specific delivery rates do not vary meaningfully as a function of geography. For this reason, the Statewide Rates shown in Table 15 were used to estimate the number of expected deliveries for individual managed care plans based on the ethnic composition of beneficiaries.

Table 16, below, shows the results of this second step. Plans are ranked based on the estimated completion rates for deliveries reported.

Table 16.
Delivery Estimates by Managed Care Plan, and
FFS Completeness Percentages

Plan Name	Delivery Estimates			FFS Completeness Percentages	
	Actual	Estimated	Actual/ Estimated (Pct)	HCP Inpat Days of FFS 99 (Pct)	HCP Inpat Days of FFS 94-95 (Pct)
Blue Cross/Kern	757	668	113%	85%	66%
Health Plan Of San Mateo	395	349	113%	120%	93%
Health Net/Fresno	589	576	102%	76%	58%
Partnership Healthplan/Solano	754	759	99%	99%	76%
Blue Cross/Sacramento	1,224	1,275	96%	76%	59%
Central Coast Alliance/Santa Cruz	312	329	95%	139%	107%
Blue Cross/San Joaquin	321	343	94%	84%	65%
Blue Cross/Fresno	2,332	2,549	91%	72%	55%
Kern Health Systems	1,078	1,213	89%	65%	50%
Blue Cross/Stanislaus	575	643	89%	74%	57%
Santa Barbara Health Initiative	565	651	87%	98%	75%
Blue Cross/Sacramento	492	583	84%	66%	51%
Blue Cross/San Diego	202	253	80%	72%	56%
Western Hlth Advantage/Sacto	229	357	64%	62%	48%
Blue Cross/Alameda	498	861	58%	76%	59%
Health Plan Of San Joaquin	717	1,263	57%	67%	52%
Contra Costa Health Plan	590	1,061	56%	57%	44%
Blue Cross/Contra Costa	79	151	52%	58%	45%
Alameda Alliance For Health	982	2,073	47%	60%	47%
Blue Cross/San Francisco	176	371	47%	51%	39%
Molina/San Bernardino	215	465	46%	37%	29%
Average:			**45%**		
CalOptima/Orange	1,230	2,859	43%	59%	46%
Maxicare/Sacramento	182	431	42%	27%	21%
Blue Cross/Santa Clara	287	713	40%	33%	26%
Health Net/Sacramento	212	575	37%	23%	18%
Community Hlth Grp/San Diego	591	1,747	34%	59%	46%
Molina/Riverside	124	383	32%	29%	23%
L.A. Care Health Plan	4,627	15,689	29%	28%	21%
Universal Care/San Diego	86	352	24%	23%	17%
Kaiser/Sacramento	80	461	17%	20%	15%
UC San Diego Health Plan	40	371	11%	14%	11%
Health Net/Los Angeles	788	10,417	8%	6%	5%
Santa Clara Family Health Plan	13	1,107	1%	2%	1%

Using the delivery estimates in Table 16, it can be seen that over 20% of the plans have less than 33% completeness in reporting, 45% had less than 50% completeness, and the overall completeness was only 45%. It is interesting that in most cases plans with above average rates for encounters were found to correspondingly have above average rates for reported deliveries. Three plans appear to have greater than 100% reporting. Although this could be due to events other than non-deliveries being reported as deliveries, more likely this is the result of some imprecision in the estimation process.

To determine how similar these delivery completion percentages by plan were to the overall inpatient completion percentages, a correlation coefficient was calculated between the two sets of completion rates. The resulting correlation coefficient (0.88) indicates a very high rate of similarity between the inpatient delivery plan completeness estimate and the overall inpatient completeness estimate though different estimation methodologies were used. This validation using a different methodology adds confidence in the estimates stated earlier as to the level of incompleteness of inpatient encounter data.

Consistency of Encounter Data

The concept of data "consistency" was used to address the issue of completeness of data addressed above. Incomplete reporting could occur on a continuous basis, from day to day or week to week due to the fact that not all services are being collected and documented at the providers office. Another potential reason for incomplete reporting is that encounter data are not reported or are reported at reduced levels for several days or weeks at a time. It must be acknowledged that factors such as varying enrollment levels, seasonality of illnesses (e.g., "the flu season"), short months (e.g., February), and number of weekends in a month may also cause some fluctuations in medical services provided.

To determine the approximate extent to which temporal gaps in encounter data reporting may contribute to incompleteness of data, the relationship between consistency of encounter data reporting by month of service, and the relative "completeness" in terms of encounter volumes by plan was examined. For this analysis, the encounter data were first broken out by month of service, plan and claim type. To derive the "consistency percentage," the three months with the lowest volumes of encounter data (lowest quartile) were compared to the three highest months (highest quartile), and a percent derived. For instance, if a plan submitted 300 encounters per service month for 3 of the 12 months in CY99, 400 encounters per month for six months, and 500 encounters for three months, the derived "consistency percentage" would be 60% [900 (i.e., 300 times 3) divided by 1500 (i.e., 500 times 3)]. The higher this consistency percentage, the less variation in the volume of reported encounter data occurred throughout the year. The same methodology was applied to the FFS data for FY94-95. The average consistency factors for the managed care plans and FFS, broken out by claim type and overall, are shown in Table 17.

Table 17.
Claims/Encounters Consistency Percentages

	Managed Care	FFS for FY94-95
Drug	48.9%	83.4%
Outpatient	53.0%	87.9%
Medical	56.9%	85.0%
Inpatient	45.3%	90.9%
Average	51.0%	86.8%

To determine if there is a relationship between these consistency percentages for managed care and the completion factors derived for encounter days, correlation coefficients for each claim type were calculated. The first was a correlation between the consistency percentage (Table 17) and encounter utilization (Table 13). The second was a correlation of the rank of each plan by claim type of their consistency factor score against all plans (1 = most consistent), and the rank of each plan by claim type of their encounter day utilization (1 = most consistent). The resulting correlation coefficients are displayed in Table 18.

Table 18.
Correlation of Consistency Percentage to Utilization Days
By Claim Type

	Consistency Percent to Utilization Days	Consistency Rank to Utilization Rank
Drug	0.65	0.40
Outpatient	0.61	0.71
Medical	0.78	0.68
Inpatient	0.74	0.67
Overall	0.77	0.75

These correlation coefficients show that there is a strong positive relationship between a high volume of submitted encounters and the consistency of submission rates over time by the health plan. Conversely, incomplete data are related to low consistency of reporting for each health plan. This suggests that consistency of data submission might be used as a proxy for data completeness in planning and carrying out research, as well as a strategy for improving data quality for DHS and the health plans.

IV. References

[1] Garnick DW, Hendricks AM, Comstock CB. Measuring quality of care: Fundamental information for administrative datasets. *Inter J Qual Hlth Care*. 1994;6:163-77.

[2] Welch WP, Welch HG. Medicare analysis. *Health Affairs* 1995;14(4):105-16.

[3] Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems. *Ann Intern Med* 1993; 119:844-9.

[4] Poses RM, Wally WR, McClish DK, Anthony M. Controlling for confounding by indication for treatment: Are administrative data equivalent to clinical data? *Med Care* 1995;33:AS36-46.

[5] Roos LL, Sharp SM, Cohen MM. Comparing clinical information with claims data: Some similarities and differences. *J Clin Epidemiol* 1991;44:881-8.

[6] Smith WR, Cotter JJ, Rossiter LF. System change: quality assessment and improvement for Medicaid managed care. *Health Care Financing Rev* 1996; 17:97-115.

[7] Schwartz RM, Gagnon DE, Muri JH, Zhao QR, Kellogg R. Section 2: Measurement -- Administrative data for quality improvement. *Pediatrics* 1999;103:291-301.

[8] Weiner JP, Powe NR, Steinwachs DM, Dent G. 1990. Applying insurance claims data to assess quality of care: A compilation of potential indicators. *Qual Rev Bull* 1990;16:424-38.

[9] McGlynn EA, Damberg CL, Kerr EA, Brook RH. Health information systems: design and analytic applications. Santa Monica: RAND Health, 1998.

[10] Kashner TM. Agreement between administrative files and written medical records. *Med Care* 1998; 36:1324-36.

[11] Rawson NSB, D'Arcy C. Assessing the validity of diagnostic information in administrative health care utilization data: Experience in Saskatchewan. *Pharmacoepidemiology and Drug Safety*. 1998;7:389-98.

[12] Fisher ES, Whaley FS, Krushat WM, Malenka DJ, Fleming C, Baron JA, Hsia DC. 1992. The accuracy of Medicare's hospital claims data: Progress has been made, but problems remain." *Am J Pub Hlth* 82(February): 243-8.

[13] Dresser MVB, Feingold L, Rosenkranz SL, Coltin KL. 1997. Clinical quality measurement: Comparing chart review and automated methodologies. *Med Care*; 35:539-52.

[14] Retchin SM, Ballard DJ. Commentary: Establishing standards for the utility of administrative claims data. *Hlth Svs Research* 1998;32:861-6.

[15] Luck J, Peabody JW, Dresselhaus TR, Lee M, Lee M, Glassman P. How well does chart abstraction measure quality? A prospective comparison of standardized patients with the medical record. *Am J Med*. 2000;108:642-9.

[16] Iezzoni LI. Assessing quality using administrative data. *Ann Int Med*. 1997,127:666-74.

[17] Palmer RH. Process-based measures of quality: The need for detailed clinical data in large health care databases. *Ann Intern Med* 1997;127:733-8.

[18] Leatherman S, Peterson E, Heinen L, Quam L. Quality screening and management using claims data in a managed care setting. *Qual Rev Bull* 1991;17:349-59.

[19] Laguna Associates. Collection and use of encounter data in a capitated managed care program. Dec 1995.

[20] Ray WA. Policy and program analysis: Using administrative databases. *Ann Intern Med* 1997;127:712-8.

[21] The MEDSTAT Group. A guide for states to assist in the collection and analysis of Medicaid managed care data, Second Edition. 1999.

V. Additional Sources

Cotter J, Smith W, Rossiter L, Pugh C, Bramble J. Combining state administrative databases and provider records to assess the quality of care for children enrolled in Medicaid. *Am J Med* 1999;14:98-104.

Diers D, Bozzo J. Using administrative data for practice and management. *Nurs Econ* 1999;17:233-7.

Editors, A framework for capturing clinical data sets from computerized sources. *Ann Intern Med* 1997;127:675-82.

Editors, Coming to grips with large databases. *Ann Int Med*. 1997,127:645-7.

Editors, High-quality clinical databases: breaking down barriers. *Lancet* 1999;353:1205-6.

Editors, The computer-based clinical record -- where do we stand? *Ann Intern Med* 1993;119:1046-8.

Elixhauser A, Steiner C, Harris DR, and Coffey R. Comorbidity measures for use with administrative data. *Med Care* 1998; 36:8-27.

Fong D, Seddon J. Use of insurance claims database to evaluate the outcomes of ophtahalmic surgery. *Surv Ophthalmol* 1997;42:271-8.

Fowles JB, Lawthers AG, Weiner JP, Garnick DW, Petrie DS, Palmer RH. Agreement between physicians' office records and Medicare part B claims data. *Health Care Financ Rev*. 1995;16:189-99.

Fox M, Phua K. Using Medicaid claims data to evaluate a large physician fee increase. *Hlth Svs Research* 1994;29:315-340.

Hannan EL, Kilburn H Jr, Lindsey ML, Lewis R. Clinical versus administrative data bases for CABG surgery. Does it matter? *Med Care*. 1992;30:892-907.

Hsia DC, Krushat WM, Fagan AB, Tebbutt JA, Kusserow RP. Accuracy of diagnostic coding for Medicare patients under prospective-payment system. *N Engl J Med* 1988; 318:352-5.

Hunt JP, Baker CC, Fakhry SM, Rutledge RR, Ransohoff D, Meyer AA. Accuracy of administrative data in trauma. *Surgery* 1999:126;191-7.

Hurley RE, Freund DS, and Paul JE, Managed Care in Medicaid (Ann Arbor: Health Administration Press, 1993).

Iezzoni LI, Foley SM, Daley J, Hughes J, Fisher ES, Heeren T. Comorbidities, complications, and coding bias. Does the number of diagnosis codes matter in predicting in-hospital mortality? *JAMA*. 1992,267:2197-03.

Kaye N, Pernice C, Pelletier H (Ed.). Medicaid managed care: A guide for states, 4th Ed. Vol. I – Part 1, Medicaid managed care program, scope and operations. National Academy for State Health Policy. March 1999.

Lange L, Jacox A. Using large data bases in nursing and health policy research. *Journal of Professional Nursing* 1993;9:204-11.

Lei J, Duisterhout JS, Westerhof HP, van der Does E, Cromme PVM, Boon WM, van Bemmel JH. The introduction of computer-based patient records in the Netherlands. *Ann Intern Med*. 1993;119:1036-41.

Mainous A, Hueston W. Using other people's data: The ins and outs of secondary data analysis. *Fam Med* 1997;29:568-71.

Mathematica. Medicaid managed care encounter data: What, why, and where next? Jan. 1996.

Office of Technology Assessment, U.S. Congress. Brining Health Care Online. The Role of Information Technologies. OTA-ITC-624. Washington, DC: US Gov Pr Off; September 1995.

Outlook Associates. Mapping the flow of eligibility and encounter data in Medi-Cal managed care. March 2001.

Powe NR, Weiner JP, Starfield B, Stuart M, Baker A, Steinwachs DM. Systemwide provider performance in a Medicaid program: profiling the care of patients with chronic illnesses. *Med Care* 1996;34:798-810.

Robinson JR, Young TK, Roos LL, Gelskey DE. Estimating the burden of disease -- Comparing administrative data and self-reports. *Med Care* 1997;35:932-947.

Romano PS. Can administrative data be used to compare the quality of health care? *Med Care Rev* 1993;50:451-77.

Roos L, Sharp S, Wajda A. Assessing data quality: A computerized approach *Soc Sci Med* 1989;28:175-182.

Roos LL, Mustard CA, Nicol JP, et al. Registries and administrative data: Organization and accuracy. *Med Care* 1993;31:201-12.

Steinwachs DM, Stuart ME, Scholle S, Starfield B, Fox MH, Weinter JP. 1998. A comparison of ambulatory Medicaid claims to medical records: A reliability assessment. *Am J Med Qual* 13: 63-9.

Tape T, Campbell J. Computerized medical records and preventive health care: Success depends on many factors. *Am J Med* 1993;94:619-5.

Udder D, Martin P. Core Data Set -- Importance to health services research, outcomes, research, and policy research *Compute Nurs*. 1997;15:538-542.

William M. Mercer, Inc., Joint Purchasers' Risk Assessment Project: Phase I (Sep 9, 2000).

Wingert, TD, Kralewski JE, Lindquist TJ, Knutson DJ. Constructing episodes of care from encounter and claims data: some methodological issues. *Inquiry* 1995/96;32:430-43.